



Unpacking the Unpredictable

From Machine Learning to Generative AI for Detecting Crises in Presidential
Democracies

Dr. Bastián González-Bustamante

Associate Professor, Universidad Diego Portales
Lecturer and Postdoctoral Researcher, Leiden University

Presentation at the PoliCICS Workshop, Universidad del Desarrollo
October 9, 2025

From Machine Learning...

Machine Learning and Political Events

DPhil Methodological Paper



Report

Machine Learning and Political Events: Application of a Semi-supervised Approach to Produce a Dataset on Presidential Cabinets

Social Science Computer Review
2023, Vol. 30, No. 14
© The Author(s) 2023
 Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/08882202231315917
journals.sagepub.com/home/sscr



Bastián González-Bustamante^{1,2}

Abstract

This paper describes the creation of a novel dataset on ministerial turnover and resignation calls in 12 presidential cabinets in Latin America from the mid-1970s to the early 2020s. The indicators on resignation calls and reallocations of cabinet members are entirely novel. Both constitute a relevant empirical contribution not only to the study of political dynamics in presidential systems and cabinet politics but also to public opinion and public policy topics. We focus on the creation of the dataset using optical recognition algorithms on press report archives together with machine learning models. The models permitted the training of ensemble semi-supervised classifiers over a period of almost 50 years. Subsequently, we provide a number of measurement validity checks to cross-validate the dataset by comparing it with similar existing data and an exploratory analysis.

Keywords

comparative politics, executives, cabinets, computational social science, machine learning

Introduction

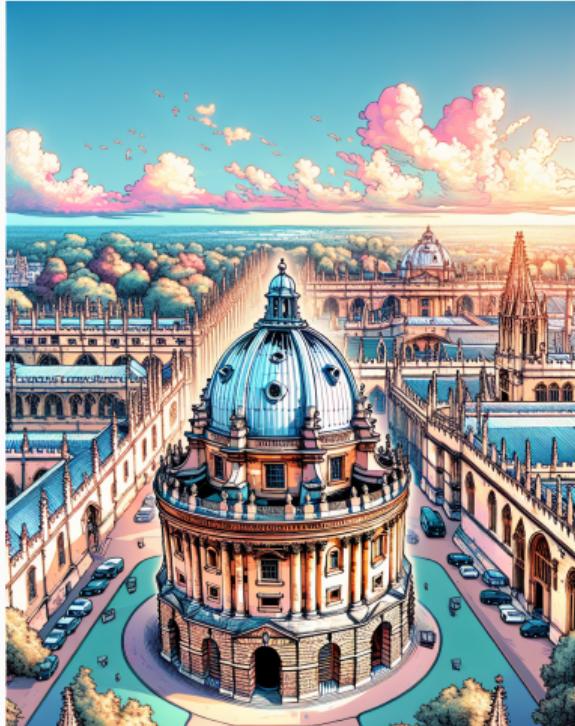
The scarcity of specific information on political events in some regions can be partly explained by the fact that, although techniques now exist for the automatic production of event data in political science, manual coding remains a common practice. Manual data gathering is costly in terms of time and resources and cannot ensure the reproducibility of the process (Eshima et al., 2021; Radford, 2019). However, since the mid-1990s, there have been growing efforts to use automated coding to produce datasets of events and phenomena relevant to political science and public affairs (e.g. Boschee et al., 2015; Schrot, 1998, 2011; Schrot & Bracke, 2013).

¹Institute of Public Administration, Leiden University, The Netherlands

²School of Public Administration, Universidad Diego Portales, Chile

Corresponding Author:

Bastián González-Bustamante, Institute of Public Administration, Faculty of Governance and Global Affairs, Leiden University, Wijnhaven, Turfmarkt 99, The Hague 2511 DP, The Netherlands.
Email: b.gonzalez.bustamante@sgg.leidenuniv.nl; bastian.gonzalez.b@msn.com



Artwork by DALL-E 3 model

Cases and Data Collection

I combined the Tesseract Optical Character Recognition (**OCR**) with different **semi-supervised machine learning models** to distinguish between resignation calls and other mentions of cabinet members in the press. This was applied to **46 years** of **Latin American Weekly Report (LAWR) archives**.



Detail of the Batches of LAWR Archives

Batch	Archives	Years	Format	Volume
Batch-01	LAWR-1975 to LAWR-1979 Including 1997	5	 CSV UTF-8	9.56 MB
Batch-02	LAWR-1980 to LAWR-1998 Excluding 1997 Including the first half of 2003	18.5	 PNG images	38.5 GB
Batch-03	LAWR-1999 to LAWR-2021 Excluding the first half of 2003	22.5	 PDF files	341 MB

Binarisation Techniques for Enhancing OCR

POSTSCRIPT					
POLITICS			PERUVIAN VIOLENCE		
MEDEGARIA GARNET Former attorney Ernesto Medegaria Garnet, 61, died Sunday. He was a former ambassador to Brazil, Argentina, Colombia and Venezuela. He also served as a representative of the United Nations in Geneva, and as a member of the International Commission of Jurists.			PERUVIAN VIOLENCE		
REGULAR SABINE Alessandro Vito, 40, of Peru's Virgin Gothic Theater and theater director Renato Pachano Lima, 40, and Hector "Rufino" Domingos Moraes Lima, 39, were killed Saturday. Rodrigo Alvarado Lima, 30, was injured.			Death toll		
Theater owner Celso Humedo and actress Ana Machado Arantes were also killed. Other theatergoers injured included Sergio Lira, 29, and actress Paula de Oliveira, 31, who was pregnant.			Wk	Mo	Tues
In addition to theatergoers killed Saturday, two others died Sunday. They were Emilio Samaniego Lima, 39, and Oscar de la Torre, 31.			1987	1,000	607
In all, 1987 has seen 2,600 deaths, 1,986 of them in August. That compares with 2,500 deaths in August 1986.			1986	2,000	1,986
Theatergoers killed Saturday included: - Alessandro Vito - Renato Pachano - Hector "Rufino" - Domingos Moraes - Celso Humedo - Ana Machado Arantes - Sergio Lira - Paula de Oliveira - Emilio Samaniego - Oscar de la Torre			1985	2,000	2,452
Deaths, 1986 and 1987, include those from August 1986 through August 1987.			1984	1,944	1,947
Deaths, 1987 through August 1987:			1983	1,739	2,393
Deaths, 1986 through August 1986:			1982	1,739	250
Deaths, 1985 through August 1985:			1981	1,692	1,692
Deaths, 1980 through August 1980:			1980	1,607	220
Deaths, 1979 through August 1979:			1979	1,500	100
Deaths, 1978 through August 1978:			1978	1,400	100
Deaths, 1977 through August 1977:			1977	1,300	100
Deaths, 1976 through August 1976:			1976	1,200	100
Deaths, 1975 through August 1975:			1975	1,100	100
Deaths, 1974 through August 1974:			1974	1,000	100
Deaths, 1973 through August 1973:			1973	900	100
Deaths, 1972 through August 1972:			1972	800	100
Deaths, 1971 through August 1971:			1971	700	100
Deaths, 1970 through August 1970:			1970	600	100
Deaths, 1969 through August 1969:			1969	500	100
Deaths, 1968 through August 1968:			1968	400	100
Deaths, 1967 through August 1967:			1967	300	100
Deaths, 1966 through August 1966:			1966	200	100
Deaths, 1965 through August 1965:			1965	100	100
Deaths, 1964 through August 1964:			1964	100	100
Deaths, 1963 through August 1963:			1963	100	100
Deaths, 1962 through August 1962:			1962	100	100
Deaths, 1961 through August 1961:			1961	100	100
Deaths, 1960 through August 1960:			1960	100	100
Deaths, 1959 through August 1959:			1959	100	100
Deaths, 1958 through August 1958:			1958	100	100
Deaths, 1957 through August 1957:			1957	100	100
Deaths, 1956 through August 1956:			1956	100	100
Deaths, 1955 through August 1955:			1955	100	100
Deaths, 1954 through August 1954:			1954	100	100
Deaths, 1953 through August 1953:			1953	100	100
Deaths, 1952 through August 1952:			1952	100	100
Deaths, 1951 through August 1951:			1951	100	100
Deaths, 1950 through August 1950:			1950	100	100
Deaths, 1949 through August 1949:			1949	100	100
Deaths, 1948 through August 1948:			1948	100	100
Deaths, 1947 through August 1947:			1947	100	100
Deaths, 1946 through August 1946:			1946	100	100
Deaths, 1945 through August 1945:			1945	100	100
Deaths, 1944 through August 1944:			1944	100	100
Deaths, 1943 through August 1943:			1943	100	100
Deaths, 1942 through August 1942:			1942	100	100
Deaths, 1941 through August 1941:			1941	100	100
Deaths, 1940 through August 1940:			1940	100	100
Deaths, 1939 through August 1939:			1939	100	100
Deaths, 1938 through August 1938:			1938	100	100
Deaths, 1937 through August 1937:			1937	100	100
Deaths, 1936 through August 1936:			1936	100	100
Deaths, 1935 through August 1935:			1935	100	100
Deaths, 1934 through August 1934:			1934	100	100
Deaths, 1933 through August 1933:			1933	100	100
Deaths, 1932 through August 1932:			1932	100	100
Deaths, 1931 through August 1931:			1931	100	100
Deaths, 1930 through August 1930:			1930	100	100
Deaths, 1929 through August 1929:			1929	100	100
Deaths, 1928 through August 1928:			1928	100	100
Deaths, 1927 through August 1927:			1927	100	100
Deaths, 1926 through August 1926:			1926	100	100
Deaths, 1925 through August 1925:			1925	100	100
Deaths, 1924 through August 1924:			1924	100	100
Deaths, 1923 through August 1923:			1923	100	100
Deaths, 1922 through August 1922:			1922	100	100
Deaths, 1921 through August 1921:			1921	100	100
Deaths, 1920 through August 1920:			1920	100	100
Deaths, 1919 through August 1919:			1919	100	100
Deaths, 1918 through August 1918:			1918	100	100
Deaths, 1917 through August 1917:			1917	100	100
Deaths, 1916 through August 1916:			1916	100	100
Deaths, 1915 through August 1915:			1915	100	100
Deaths, 1914 through August 1914:			1914	100	100
Deaths, 1913 through August 1913:			1913	100	100
Deaths, 1912 through August 1912:			1912	100	100
Deaths, 1911 through August 1911:			1911	100	100
Deaths, 1910 through August 1910:			1910	100	100
Deaths, 1909 through August 1909:			1909	100	100
Deaths, 1908 through August 1908:			1908	100	100
Deaths, 1907 through August 1907:			1907	100	100
Deaths, 1906 through August 1906:			1906	100	100
Deaths, 1905 through August 1905:			1905	100	100
Deaths, 1904 through August 1904:			1904	100	100
Deaths, 1903 through August 1903:			1903	100	100
Deaths, 1902 through August 1902:			1902	100	100
Deaths, 1901 through August 1901:			1901	100	100
Deaths, 1900 through August 1900:			1900	100	100
Deaths, 1899 through August 1899:			1899	100	100
Deaths, 1898 through August 1898:			1898	100	100
Deaths, 1897 through August 1897:			1897	100	100
Deaths, 1896 through August 1896:			1896	100	100
Deaths, 1895 through August 1895:			1895	100	100
Deaths, 1894 through August 1894:			1894	100	100
Deaths, 1893 through August 1893:			1893	100	100
Deaths, 1892 through August 1892:			1892	100	100
Deaths, 1891 through August 1891:			1891	100	100
Deaths, 1890 through August 1890:			1890	100	100
Deaths, 1889 through August 1889:			1889	100	100
Deaths, 1888 through August 1888:			1888	100	100
Deaths, 1887 through August 1887:			1887	100	100
Deaths, 1886 through August 1886:			1886	100	100
Deaths, 1885 through August 1885:			1885	100	100
Deaths, 1884 through August 1884:			1884	100	100
Deaths, 1883 through August 1883:			1883	100	100
Deaths, 1882 through August 1882:			1882	100	100
Deaths, 1881 through August 1881:			1881	100	100
Deaths, 1880 through August 1880:			1880	100	100
Deaths, 1879 through August 1879:			1879	100	100
Deaths, 1878 through August 1878:			1878	100	100
Deaths, 1877 through August 1877:			1877	100	100
Deaths, 1876 through August 1876:			1876	100	100
Deaths, 1875 through August 1875:			1875	100	100
Deaths, 1874 through August 1874:			1874	100	100
Deaths, 1873 through August 1873:			1873	100	100
Deaths, 1872 through August 1872:			1872	100	100
Deaths, 1871 through August 1871:			1871	100	100
Deaths, 1870 through August 1870:			1870	100	100
Deaths, 1869 through August 1869:			1869	100	100
Deaths, 1868 through August 1868:			1868	100	100
Deaths, 1867 through August 1867:			1867	100	100
Deaths, 1866 through August 1866:			1866	100	100
Deaths, 1865 through August 1865:			1865	100	100
Deaths, 1864 through August 1864:			1864	100	100
Deaths, 1863 through August 1863:			1863	100	100
Deaths, 1862 through August 1862:			1862	100	100
Deaths, 1861 through August 1861:			1861	100	100
Deaths, 1860 through August 1860:			1860	100	100
Deaths, 1859 through August 1859:			1859	100	100
Deaths, 1858 through August 1858:			1858	100	100
Deaths, 1857 through August 1857:			1857	100	100
Deaths, 1856 through August 1856:			1856	100	100
Deaths, 1855 through August 1855:			1855	100	100
Deaths, 1854 through August 1854:			1854	100	100
Deaths, 1853 through August 1853:			1853	100	100
Deaths, 1852 through August 1852:			1852	100	100
Deaths, 1851 through August 1851:			1851	100	100
Deaths, 1850 through August 1850:			1850	100	100
Deaths, 1849 through August 1849:			1849	100	100
Deaths, 1848 through August 1848:			1848	100	100
Deaths, 1847 through August 1847:			1847	100	100
Deaths, 1846 through August 1846:			1846	100	100
Deaths, 1845 through August 1845:			1845	100	100
Deaths, 1844 through August 1844:			1844	100	100
Deaths, 1843 through August 1843:					

Note. The images correspond to a LAWR page before and after binarisation techniques.

Tesseract Optical Character Recognition



How Domingo Cavallo rose to become Menem's virtual prime minister

Argentina are still hotly debating what lay behind the sudden resignations of President Raúl Alfonsín and his replacement, Domingo Cavallo. His rapid, uncontrolled usurpation of the role of economy minister – dismissed as suffi-

cient reason for his removal – has been published about all sorts of political plots. One that seems to stand up is Cavallo's desire to depose his mentor at Menem – and, according to sources with access to the president, one day accept it. However, a widely accepted version is that Menem, seeing Cavallo as a popular Cavallo as his economy minister, to preside over the period of transition, has given him carte blanche. It would follow a first phase of stabilisation. In this scenario, all Cavallo did was to choose his own timing.

■

INSIDE

PERU: Wildlife spread of cholera epidemic. (2)

EQUADOR: Holiday disease impact on oil. (2)

COLOMBIA: Presidents of Wta, Mirabal, with CHL. (2)

VENEZUELA: Lower prices force budget cutback. (2)

BOLIVIA: A Soviet partner for Latin America? (2)

CHILE: Proposal to extend Ayala's mandate. (6)

COSTA RICA: Preparing a new batch of cases. (11)

EL SALVADOR: Progress of talks, aimed, with CHL. (2)

Another strand of interpretation is that the change took place as a result of events he had not bargained with. Events have forced him to be able to organise the share-out of key ministerial posts and even to appoint men of his own party. Cavallo is said, however, to be organised about all sorts of political plots.

Many expect Cavallo to push for an agreement with the main opposition party, Raúl Alfonsín's Union Democratico (UDC), to form a new form of coalition that would ensure the continuity of the government. This has been rifle about which of the many constitutional succession procedures would find most favour.



How Domingo Cavallo rose to become Menem's virtual prime minister

Argentina are still hotly debating what lay behind the sudden resignations of President Raúl Alfonsín and his replacement, Domingo Cavallo (WR-01-01). The scenario that seems to stand up is Cavallo's desire to depose his mentor at Menem – and, according to sources with access to the president, one day accept it. However, a widely accepted version is that Menem, seeing Cavallo as a popular Cavallo as his economy minister, to preside over the period of transition, has given him carte blanche. It would follow a first phase of stabilisation. In this scenario, all Cavallo did was to choose his own timing.

■

INSIDE

PERU: Wildlife spread of cholera epidemic. (2)

BOLIVIA: A Soviet partner for Latin America? (2)

CHILE: Proposal to extend Ayala's mandate. (6)

COSTA RICA: Preparing a new batch of cases. (11)

EL SALVADOR: Progress of talks, aimed, with CHL. (2)

Another strand of interpretation is that the change took place as a result of events he had not bargained with. Events have forced him to be able to organise the share-out of key ministerial posts and even to appoint men of his own party. Cavallo is said, however, to be organised about all sorts of political plots.

The at a time when even such highly placed members of his government as Vice-President Eduardo Duhalde and Minister of Finance, Alfonso Cárdenas, have been instrumental in the manoeuvring to engineer Menem's departure from the Presidency. Cavallo, as economy minister, to preside over the period of transition, has given him carte blanche. It would follow a first phase of stabilisation. In this scenario, all Cavallo did was to choose his own timing.

■



How Domingo Cavallo rose to become Menem's virtual prime minister

Argentina are still hotly debating what lay behind the sudden resignations of President Raúl Alfonsín and his replacement, Domingo Cavallo (WR-01-01). The scenario that seems to stand up is Cavallo's desire to depose his mentor at Menem – and, according to sources with access to the president, one day accept it. However, a widely accepted version is that Menem, seeing Cavallo as a popular Cavallo as his economy minister, to preside over the period of transition, has given him carte blanche. It would follow a first phase of stabilisation. In this scenario, all Cavallo did was to choose his own timing.

■

INSIDE

PERU: Wildlife spread of cholera epidemic. (2)

BOLIVIA: A Soviet partner for Latin America? (2)

CHILE: Proposal to extend Ayala's mandate. (6)

COSTA RICA: Preparing a new batch of cases. (11)

EL SALVADOR: Progress of talks, aimed, with CHL. (2)

Another strand of interpretation is that the change took place as a result of events he had not bargained with. Events have forced him to be able to organise the share-out of key ministerial posts and even to appoint men of his own party. Cavallo is said, however, to be organised about all sorts of political plots.

The at a time when even such highly placed members of his government as Vice-President Eduardo Duhalde and Minister of Finance, Alfonso Cárdenas, have been instrumental in the manoeuvring to engineer Menem's departure from the Presidency. Cavallo, as economy minister, to preside over the period of transition, has given him carte blanche. It would follow a first phase of stabilisation. In this scenario, all Cavallo did was to choose his own timing.

■

It was at this point, the scenario goes, that Cavallo bolstered his position by forcing the resignation of Emilio González and his replacement by Domingo Cavallo (WR-01-01). The scenario that seems to stand up is Cavallo's desire to depose his mentor at Menem – and, according to sources with access to the president, one day accept it. However, a widely accepted version is that Menem, seeing Cavallo as a popular Cavallo as his economy minister, to preside over the period of transition, has given him carte blanche. It would follow a first phase of stabilisation. In this scenario, all Cavallo did was to choose his own timing.

■

INSIDE

PERU: Wildlife spread of cholera epidemic. (2)

BOLIVIA: A Soviet partner for Latin America? (2)

CHILE: Proposal to extend Ayala's mandate. (6)

COSTA RICA: Preparing a new batch of cases. (11)

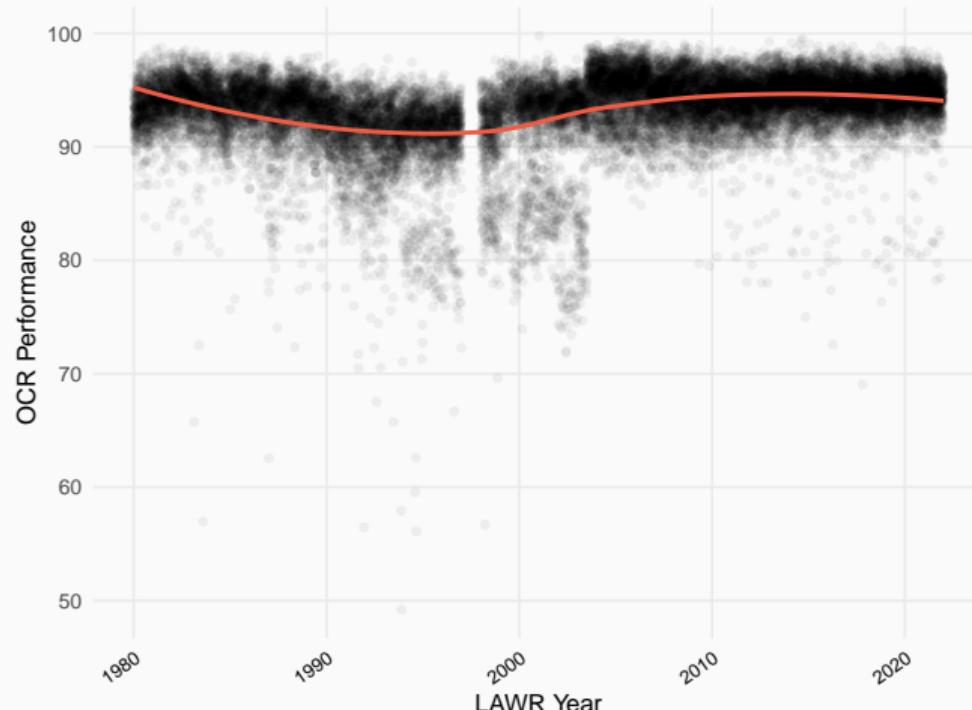
EL SALVADOR: Progress of talks, aimed, with CHL. (2)

It was at this point, the scenario goes, that Cavallo bolstered his position by forcing the resignation of Emilio González and his replacement by Domingo Cavallo (WR-01-01). The scenario that seems to stand up is Cavallo's desire to depose his mentor at Menem – and, according to sources with access to the president, one day accept it. However, a widely accepted version is that Menem, seeing Cavallo as a popular Cavallo as his economy minister, to preside over the period of transition, has given him carte blanche. It would follow a first phase of stabilisation. In this scenario, all Cavallo did was to choose his own timing.

■

Tab-Stop Lines — Column Layout — Segmented Blocks

Performance and Accuracy of the Tesseract OCR Algorithm



Accuracy is based on the proportion of text that was correctly identified.

This calculation was performed after extracting stop words and carried out with dictionaries trained with **Long Short-Term Memory (LSTM) models** used by Tesseract.

Semi-Supervised Approach

There are two significant **drawbacks** to training **supervised algorithms**:

1. This would have required a major amount of labelled text data (**time and resources**)
2. The accuracy of the algorithms is expected to decrease out-of-the-sample and over time
(batches of archives cover almost 50 years)

Therefore, I preferred a **semi-supervised approach** that could maintain a **high accuracy level over time with a small amount of coded data** for training.

1. I drew a subsample of 1,000 LAWR press releases
2. A team of human coders classified all these observations (**resignation calls**)
3. Labelling was conducted to generate ground-truth labels to cross-validate
4. Semi-supervised models using knowledge-based seed words

Semi-Supervised Approach

Pre-processing

80% of the subsample for training

20% for validation

100 tokens per news item *

SMART stop words **

Customised list ***

Stemming algorithm

* until 500 for robustness checks

** Snowball for robustness checks

*** based on LAWR numbering

Models

Naïve Bayes (NB)

Support Vector Machine (SVM) *

Radial Basis Function (RBF) SVM **

Random Forest with 100 trees (RF100)

Random Forest with 500 trees (RF500)

Extreme Gradient Boosting (XGBoost)

* Kernel-based Linear

** Kernel-based Gaussian

Labelling Process for Benchmarking Supervised Models

Coder	Labels	$M(T)$ Label	\sum Time	Consensus
1	966	39s	10h 25m	94.05
2	718	1m 15s	15h 0m	94.84
3	704	38s	7h 29m	94.82
4	318	1m 4s	5h 38m	95.47
5	263	17s	1h 14m	99.26
6	101	1m 17s	2h 9m	88.50

Note. Each observation was hand-coded by three different coders. 2.3% of observations were revised and relabelled during an iteration step.

Goodness-of-Prediction of the Semi-Supervised Models

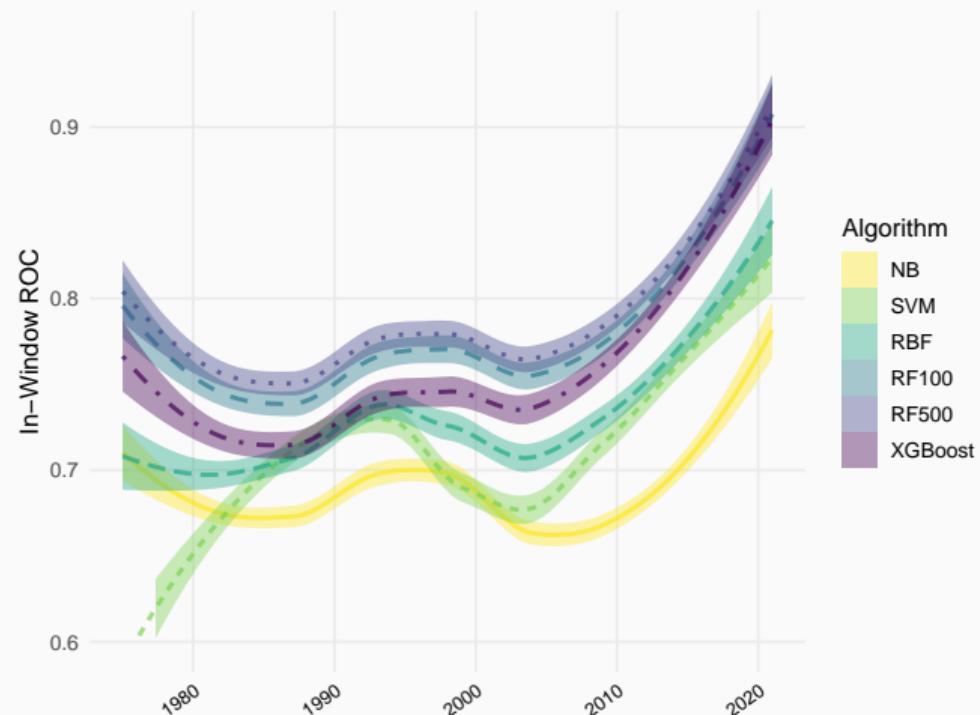
Model	Knowledge-Based Seed Words				
	Accuracy	Precision	Recall	F1-score	ROC
Naïve Bayes	0.713	0.734	0.948	0.828	0.638
Kernel Linear SVM	0.698	0.741	0.900	0.812	0.614
Kernel Gaussian RBF SVM	0.726	0.728	0.995	0.840	0.647
Random Forest (100 trees)	0.743	0.742	0.991	0.849	0.671
Random Forest (500 trees)	0.743	0.740	0.997	0.849	0.682
XGBoost	0.726	0.762	0.905	0.827	0.660

Note. The evaluation was performed with resampling methods using the testing set and 10-fold cross-validation.

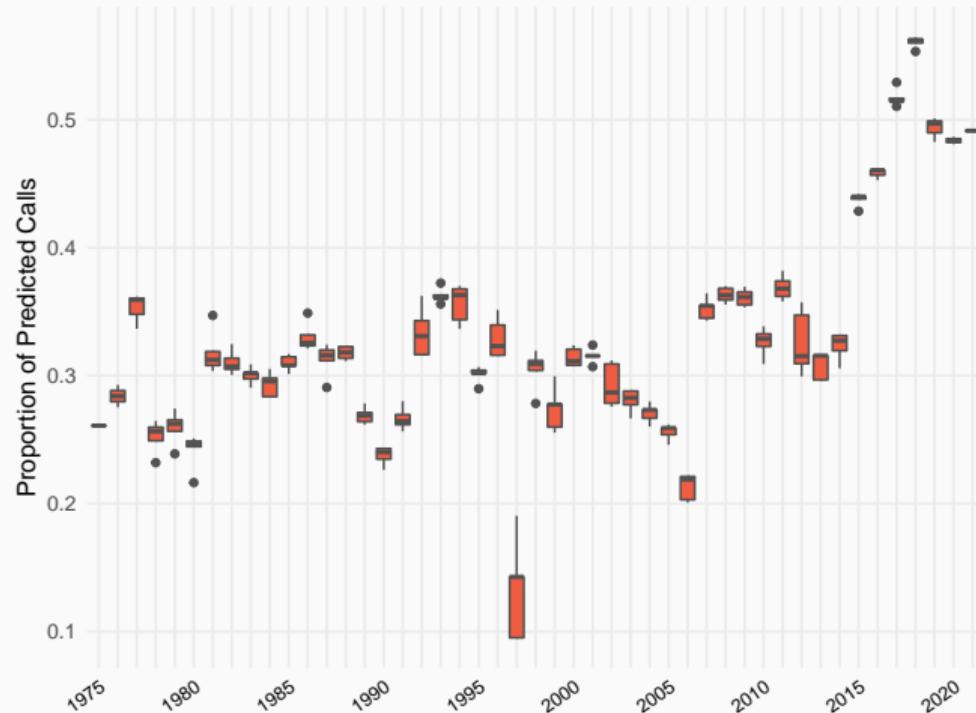
Goodness-of-Prediction of Semi-Supervised Models (1975-2021)

I trained semi-supervised models using a **five-year fixed rolling window** from 1975 to 2021 to train algorithms and predict resignation calls.

Random Forest classifiers with 500 trees 10-fold cross-validated.



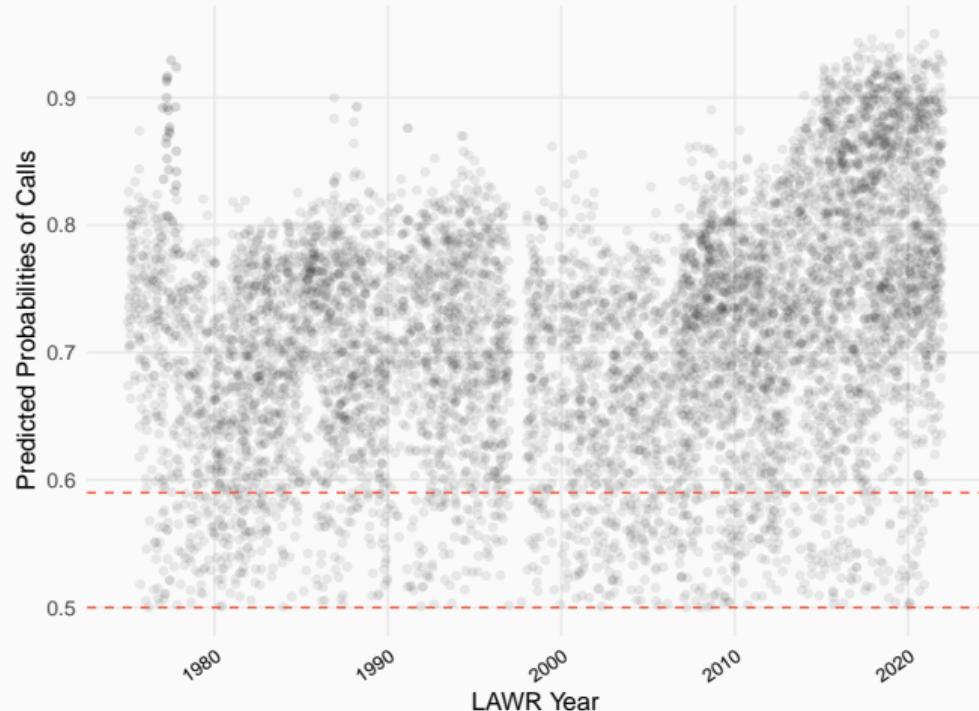
Predicted Ministerial Resignation Calls over Cabinet Press Mentions



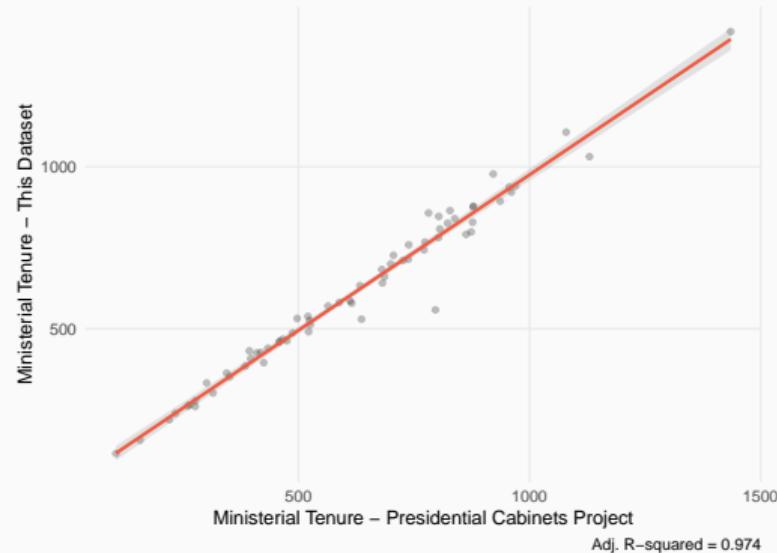
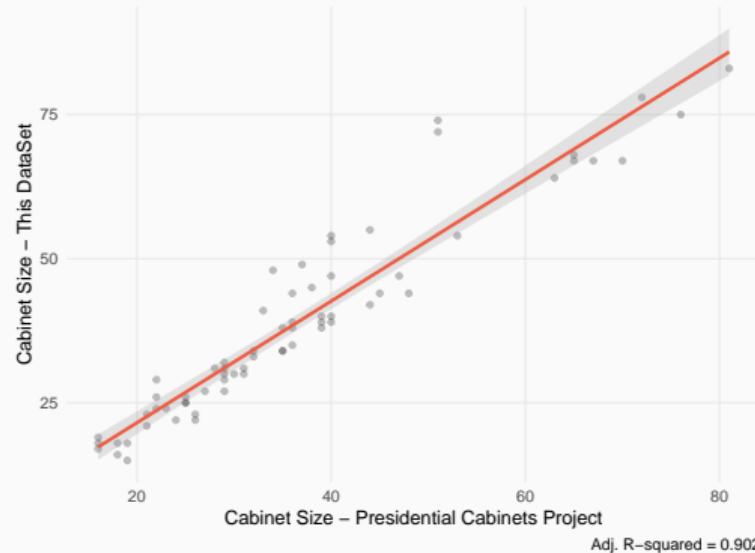
Proportion of predicted resignation calls over mentions of cabinet members in press reports, were carried out with an **ensemble semi-supervised RF500 algorithm**.

Human-in-the-Loop

I performed an augmented artificial intelligence process based on a **human review** of the predicted observations with low confidence values to improve the data's quality.

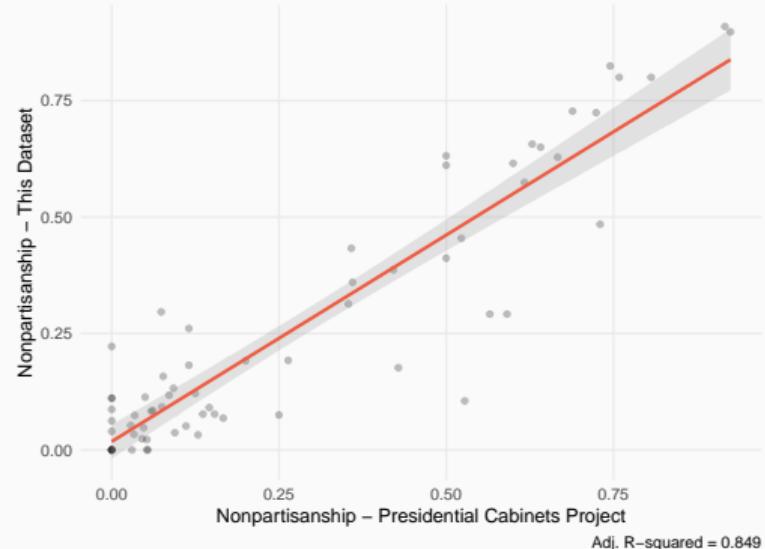
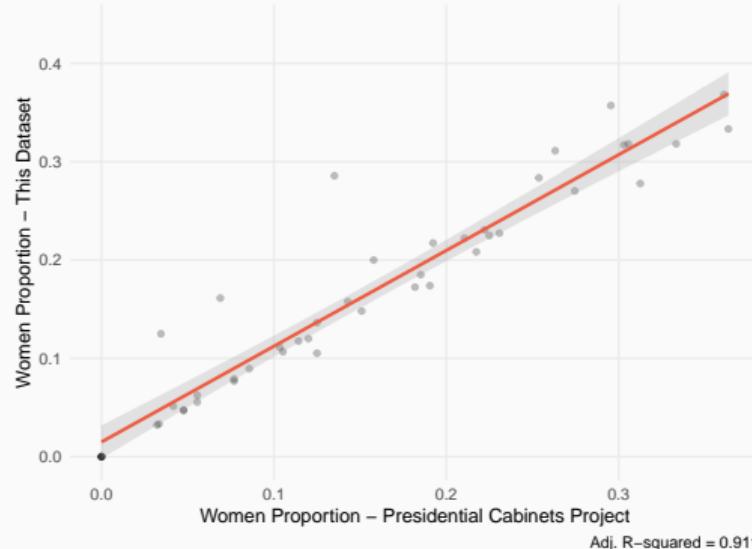


Convergent Validation against Similar Datasets



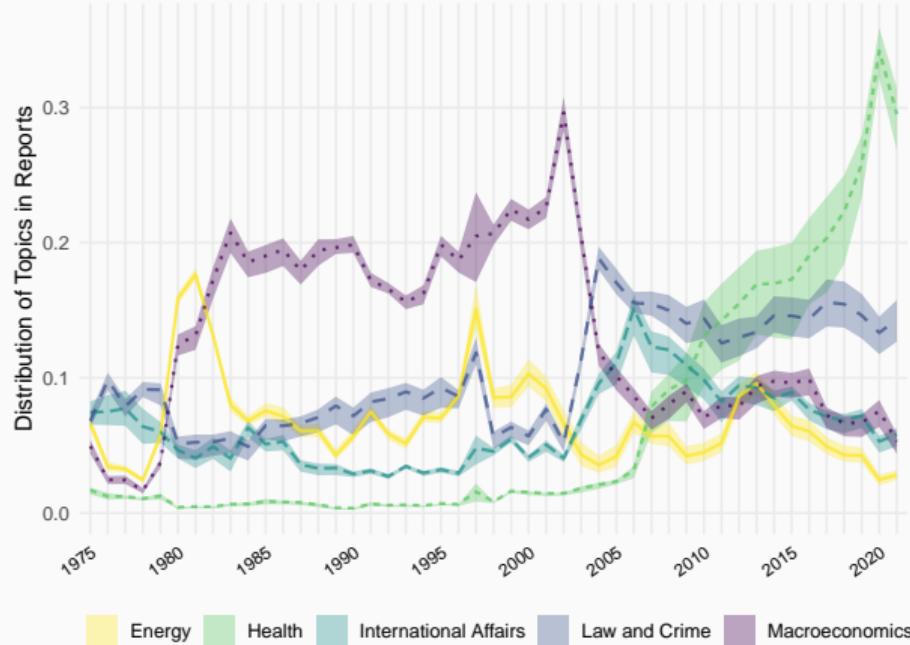
Note. Concurrent presidential terms were used for comparisons with the Presidential Cabinet Project.

Convergent Validation against Similar Datasets



Note. Concurrent presidential terms were used for comparisons with the Presidential Cabinet Project.

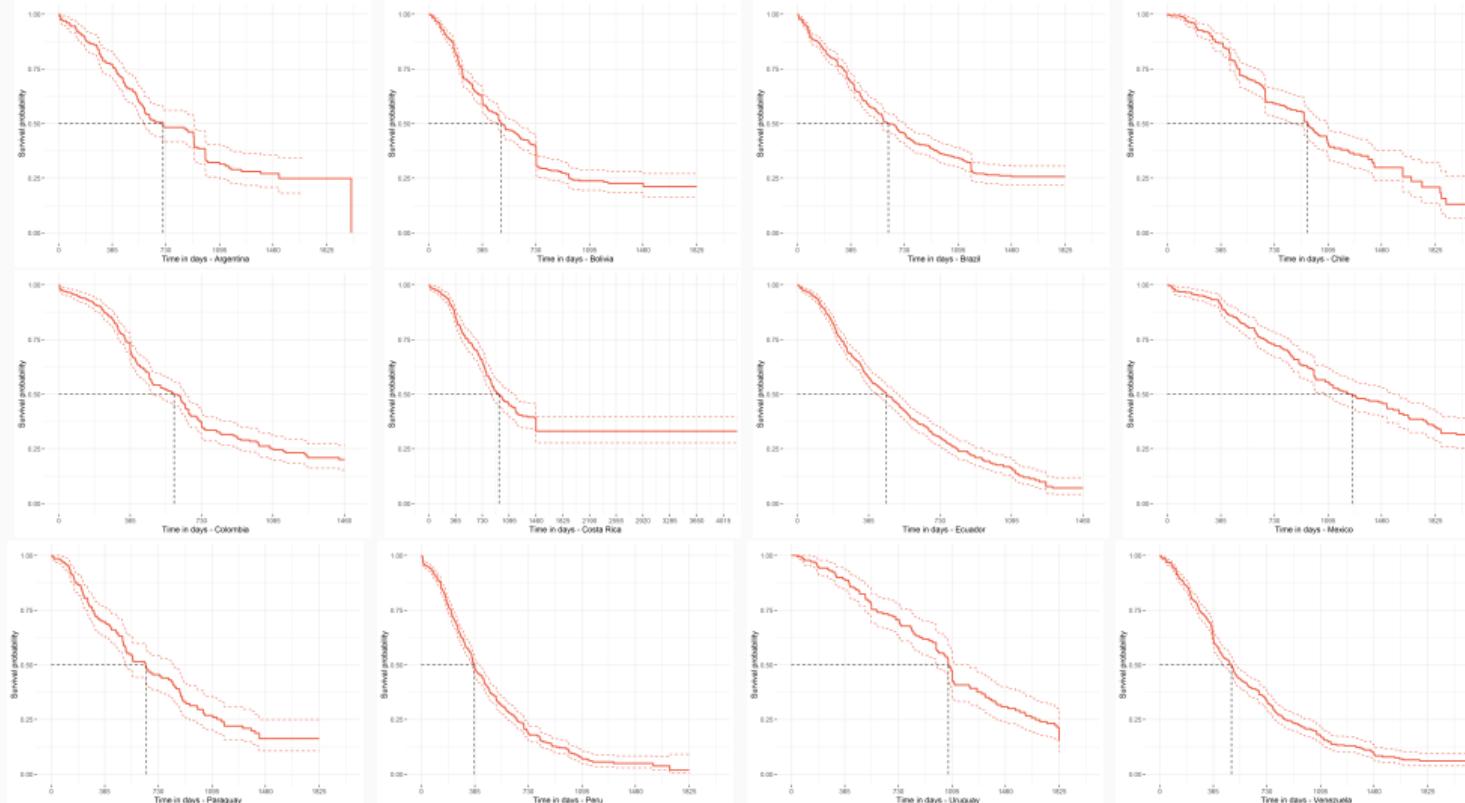
Content Validation: Major Agenda Topics in Latin America



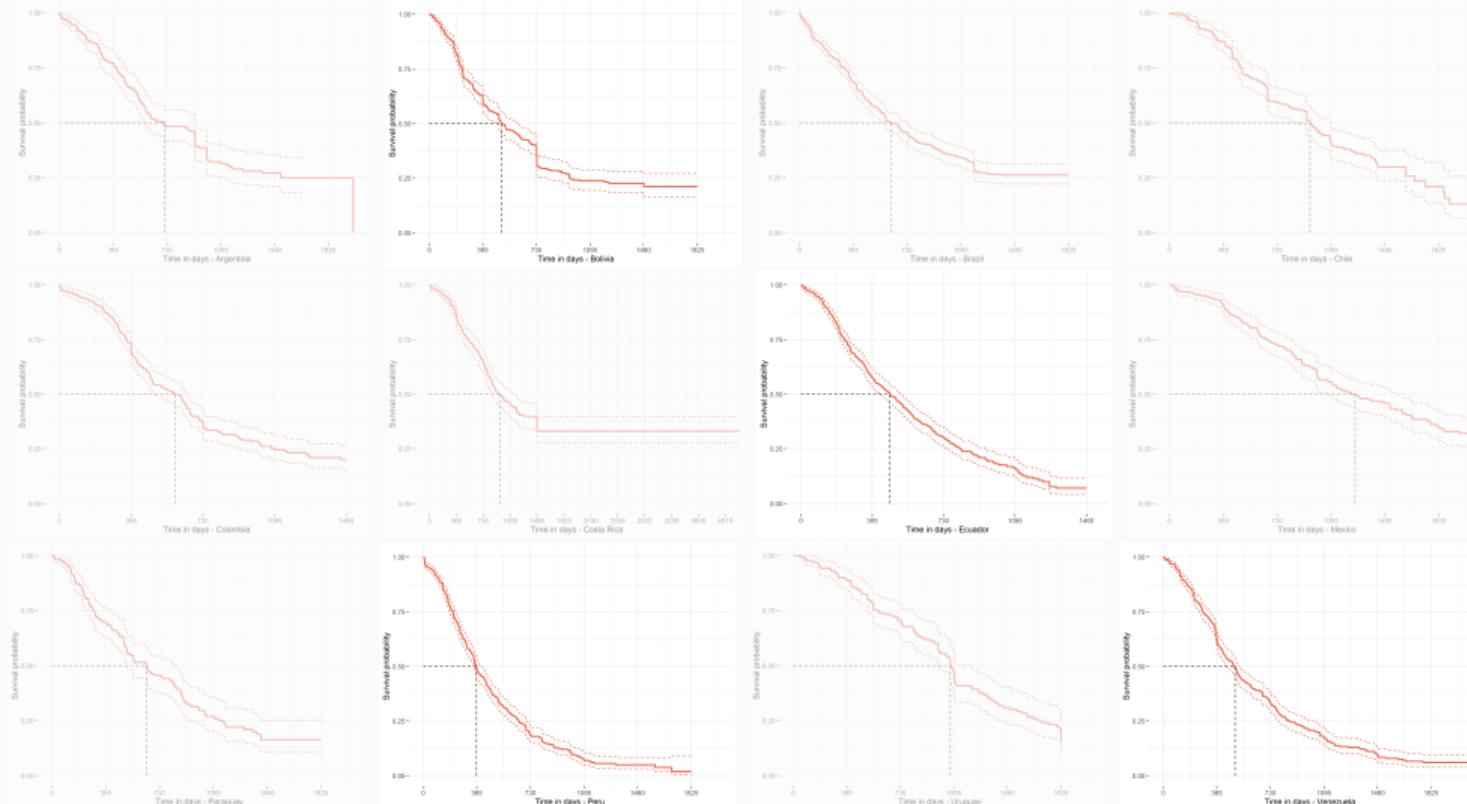
I used **Keyword Assisted Topic Modelling (key-ATM)** and a dictionary based on Comprative Agendas Project (CAP).

It is a semi-supervised Bayesian application that allows me to obtain θ (distribution of topics in the documents and Cls).

Kaplan-Meier Survival Estimations in 12 Presidential Cabinets

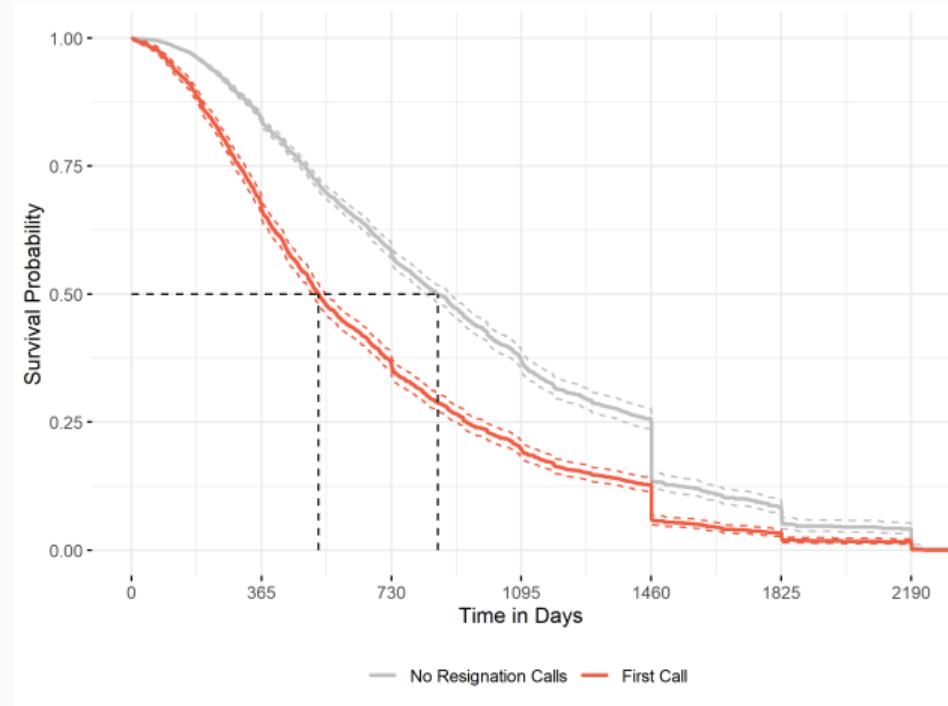


Kaplan-Meier Survival Estimations in 12 Presidential Cabinets



Kaplan-Meier Survival Estimation

$F(t) = 1 - S(t) = \Pr(T \leq t)$ can be used to obtain the incidence of ministerial exits during T by considering $S(t)$ as a survival function, if raw exit is considered an event in the **absence of competing risks**.



... to Generative AI

Enlace-Inserción UDP 2025

Unpacking the Unpredictable: Cabinet Politics in Presidential Democracies

Examining how NLP and LLMs can shed light on cabinet responses to stochastic events in Latin American democracies

Dr. Bastián González-Bustamante, Principal Investigator

Enlace-UDP Design

Project Overview



Artwork by DALL-E 3 model

- ❖ **Focus.** Investigating the impact of stochastic events on cabinet stability in presidential systems.
- ❖ **Scope.** 12 Latin American democracies, mid-1970s to early 2020s.
- ❖ **Methodology.** Utilising machine learning and AI techniques, including LLMs.

Introducing LLMs to the Field

- 📄 **Traditional text analysis.** Traditional approaches to analysing text data in social sciences often relied on manual coding, dictionaries, or topic modelling, which can be time-consuming and resource-intensive.
- 🌐 **The power of LLMs.** LLMs, with their zero- and few-shot learning capabilities, offer a powerful alternative, enabling researchers to process and analyse vast amounts of text data efficiently.
- 🏛️ **Transforming research.** This shift towards AI-powered analysis has the potential to transform computational social science research.

OpenAI's GPTs: A Closer Look

GPT and research

Researchers have increasingly utilised OpenAI's GPT models through their **API for various tasks**, including text annotation and classification.

Advantages of GPT models

GPT models offer several advantages, including **ease of use**, access to substantial computational resources, and **impressive performance** in many tasks.

Concerns and limitations

However, concerns remain regarding **reproducibility**, data privacy, and the potential for bias in these proprietary models.

The Rise of Open-Source LLMs

Addressing the limitations

Open-source LLMs have emerged as a response to concerns surrounding proprietary models, emphasising transparency, **reproducibility**, and community-driven development.

Performance and potential

Recent studies have shown that open-source LLMs can match or **even surpass** the performance of proprietary models in specific text annotation tasks, particularly in zero-shot classification.

Cost-effective alternatives

Some open-source “Small Language Models” (SLMs) offer comparable performance with significantly **reduced computational requirements** and costs.

Case Selection Strategy

Brazil

South America's largest democracy, featuring a complex federal system and a diverse political landscape.

Venezuela

Represents significant institutional changes and varying levels of democratic stability over the study period.

Costa Rica

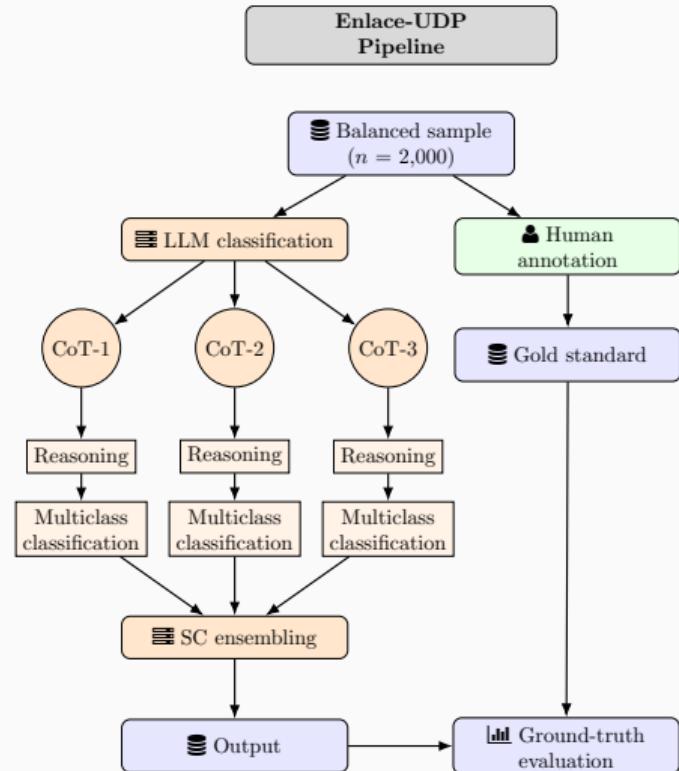
Known for consistent democratic governance and a stable institutional framework.

Mexico

Features transition from dominant-party system to competitive democracy.

Project Pipeline

- 500 media reports per country.
- 3 human coders per report.
- **Gold standard** creation and measurement validity assessment.
- We may slightly increase the sample and go for some **fine-tuned models**.



Benchmarking LLMs for Annotation

Benchmarking LLMs for Annotation

Measuring Complexity and Reproducibility: A Comprehensive Benchmark of LLMs for Multilingual Policy Agenda Topic Annotation*

Bastian González-Bustamante[†]

Leiden University

June 3, 2025

■ Please do not distribute this version without the permission of the author

Abstract

Although Large Language Models (LLMs) have rapidly expanded the text-as-data toolkit available to social scientists, their performance remains highly sensitive to task complexity, prompting, language and model parameters. In order to assess these limits, we conduct a cross-lingual benchmark of 79 contemporary LLMs (e.g., OpenAI's GPTs and o-series, Claude models, xAI's Grok, Meta's Llama, Alibaba's Qwen, Mistral models, among others) on a demanding zero-shot classification task: labelling one of the 21 major topics in the Comparative Agendas Project codebook to bills and Acts drawn from Danish, Dutch, English, French, Hungarian, Italian, Portuguese and Spanish corpora. We then conducted a meta-analysis that shows reasoning models lift F1-scores by around nine percentage points on average. In comparison, reproducible models under deterministic deployment conditions incur a cost of about eight percentage points in F1-score. In addition, preliminary and ongoing fine-tuning experiments suggest that fine-tuned transformers may surpass the best zero-shot LLM classification performance. These findings quantify the trade-off between performance and reproducibility, highlighting chain-of-thought reasoning as

Charting Reproducibility and Performance: LLMs in Multilingual Toxic Speech Detection*

Bastián González-Bustamante[†]

Leiden University

June 10, 2025

■ Please do not distribute this version without the permission of the author

Abstract

Large Language Models (LLMs) are increasingly central to analysing and mitigating incivility and toxicity in online communication, however, their comparative strengths vary by language coverage, model openness, and other factors. Drawing on the Multilingual Text Detoxification (TextDetox) corpus, which spans seven languages (i.e., Arabic, Chinese, English, German, Hindi, Russian, and Spanish), this paper benchmarks 807 model-language pairs and pools goodness-of-prediction indicators in a meta-analysis. The evaluation includes OpenAI's GPTs, o-series, Claude models, xAI's Grok, Llama checkpoints, Alibaba Qwen-series, Mistral models, among others. Three patterns emerge from our analysis. First, high-resource languages (i.e., English, German, and Spanish) enjoy, on average, a 7.7-point F1-score advantage over lower-resource counterparts (i.e., Arabic, Chinese, Hindi, and Russian). Second, proprietary models top the leaderboard in low-resource languages, however, the openness penalty is minor and statistically inconclusive. Conversely, open-source models tend to equal closed models in high-resource languages. Third, reasoning models and chain-of-thought (CoT) neither help nor harm this binary classification task, whereas compact models (with 7B parameters or fewer) trail a larger baseline by roughly 9.8 points. Ancillary findings derived from temperature

*This version was prepared for delivery at the Annual Conference of the Comparative Agendas Project (CAP), University of Konstanz, Konstanz, Germany, June 11–13, 2025.

[†]Post-doctoral Researcher in Computational Social Science and Lecturer in Governance and Development, Institute of Public Administration, Faculty of Governance and Global Affairs, Leiden University, Netherlands. ♀ Wijnhaven, Turfmarkt 99, The Hague 2511 DP, Netherlands, ☎ b.a.gonzalez.bustamante@fgga.leidenuniv.nl, ✉ <https://bgonzalezbustamante.com>, ORCID iD <https://orcid.org/0000-0003-1510-6820>.

^{*}This version was prepared for delivery at the LLM Pre-Conference Workshop, European Political Science Association (EPSA), Universidad Carlos III de Madrid, Madrid, Spain, June 25, 2025. An early version of this work was presented at the 7th Annual COMPTEXT Conference, University of Vienna, Austria, April 24–26, 2025. I am grateful to Chico Carmago for his helpful comments.

[†]Post-doctoral Researcher in Computational Social Science and Lecturer in Governance and Development, Institute of Public Administration, Faculty of Governance and Global Affairs, Leiden University, Netherlands. ♀ Wijnhaven, Turfmarkt 99, The Hague 2511 DP, Netherlands, ☎ b.a.gonzalez.bustamante@fgga.leidenuniv.nl, ✉ <https://bgonzalezbustamante.com>, ORCID iD <https://orcid.org/0000-0003-1510-6820>.

Ground-Truth Policy Agenda

Denmark

15 101 bills
1953–2016

NLD

4 684 bills
1981–2009

UK

6 169 Acts
1911–2015

France

3 069 laws
1979–2013

Hungary

8 220 bills
1990–2022

Italy

4 554 laws
1983–2013

Brazil

2 449 laws
2003–2014

Spain

2 256 laws-decrees
1980–2018

Note. We split the samples in a proportion of 70/15/15 (stratified by major topic) for training, validation, and testing for future fine-tuning jobs. The samples correspond to ground-truth data of the Comparative Agendas Project.

Ground-Truth Toxicity

Arabic 5000 tweets manually annotated	Chinese 5000 messages for toxic detection	English 5000 Wikipedia comments	German 5000 Twitter and Facebook comments
Hindi 5000 Twitter and Facebook comments	Russian 5000 comments on social network OK	Spanish 5000 messages for toxic detection	

Note. We split the samples in a proportion of 70/15/15 for training, validation, and testing in case of fine-tuning jobs. The samples correspond to ground-truth data prepared for [CLEF TextDetox \(2024\)](#).

79/109 LLMs

Multiclass/Binary

Run **620/807 times** under different conditions (e.g., parameters, API/local, temperature, datasets/language)

* GPT-5 and OSS (August 2025).



SOTA closed-source LLMs

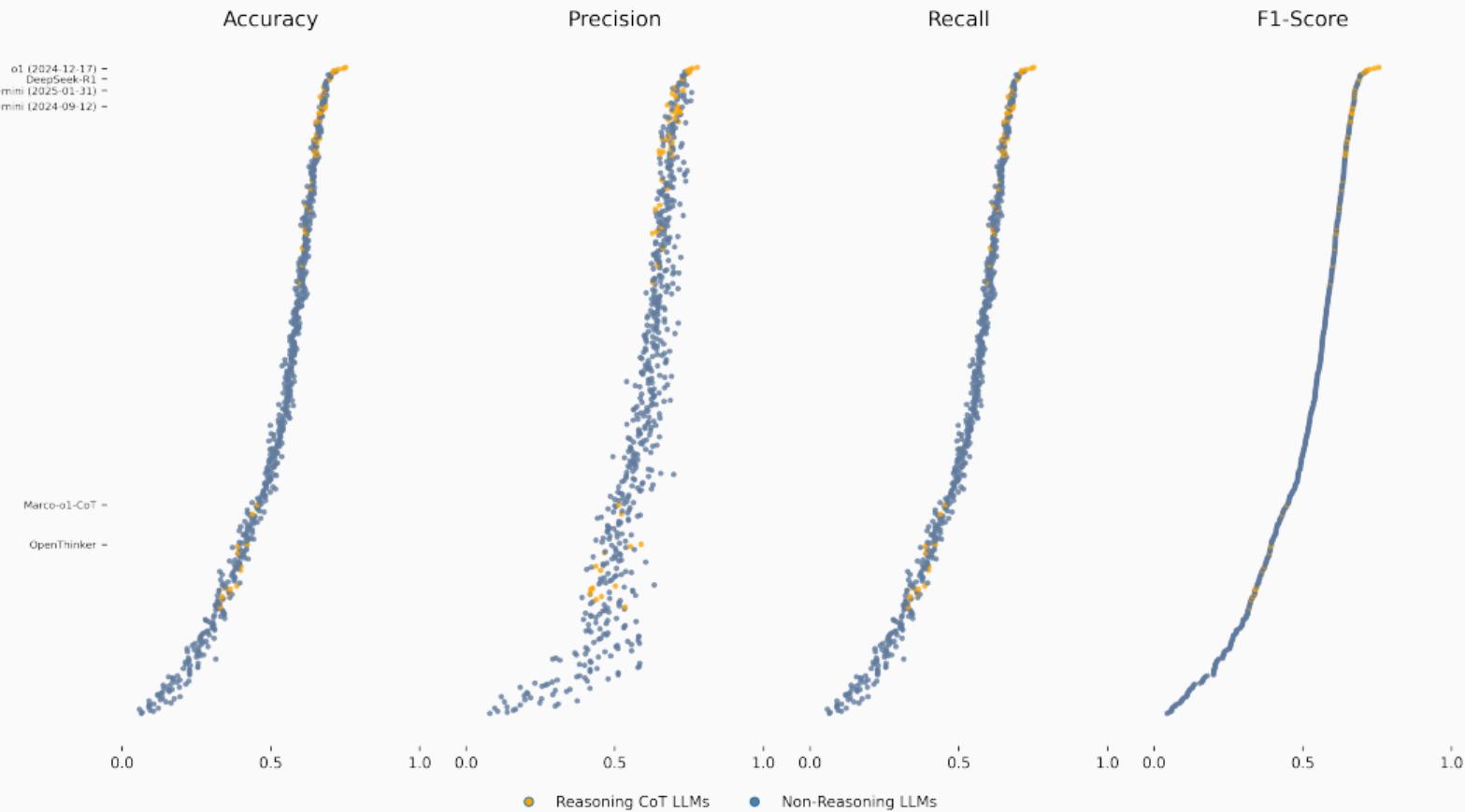
o4-mini, o3-mini, o1, GPT-4.1, GPT-4.5-preview, Grok 3 Beta, Claude 4 Opus, among others



SOTA open-source LLMs

Llama 4 Maverick (400B) and Scout (107B), Mistral 3.1 (24B), Llama 3.3 (70B), DeepSeek-R1 (671B), DeepSeek-V3 (671B), among others

Benchmarking Policy Agenda



Top Performing Models Policy Agenda

0.76

**Maximum
F1-Score**

Achieved by OpenAI
reasoning models

0.69

**Top Open-
Source Models**

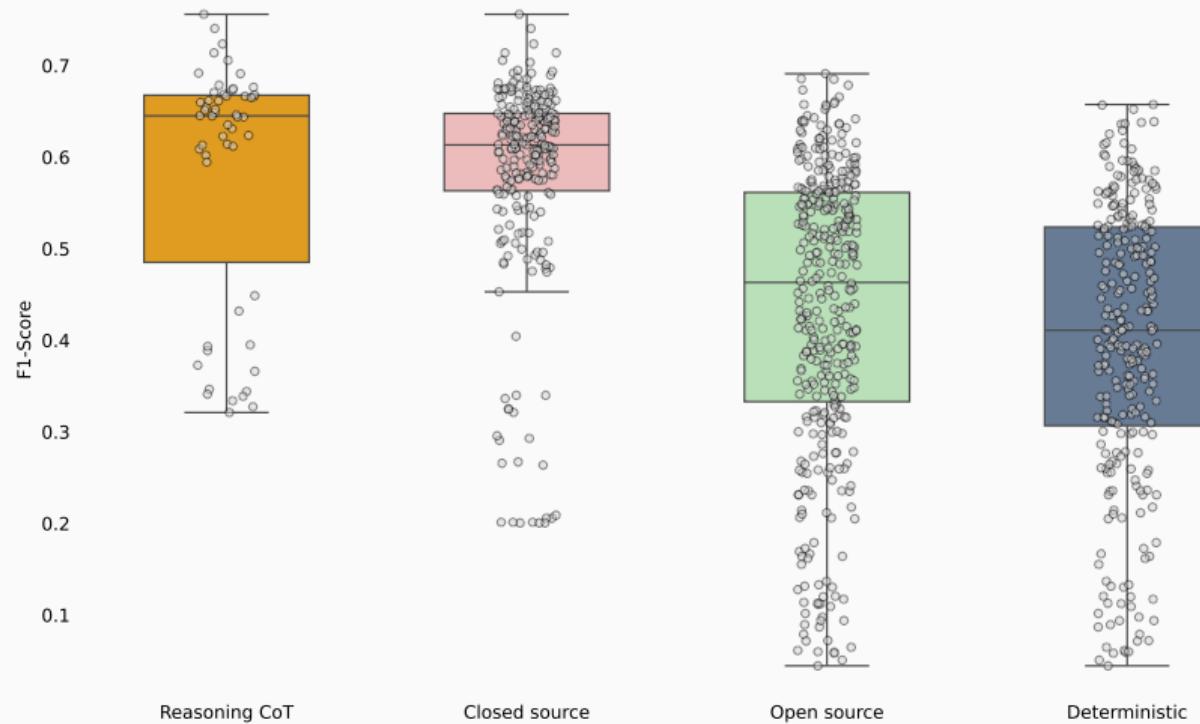
DeepSeek-R1 and V3 (671B)
and Llama 3.1 (405B)

0.54

**Mid-Range
Performance**

Gemini 1.5 series,
GPT4o-mini, Mis-
tral models, Qwen 2.5

F1-Score Across Reasoning, Source and Deterministic Setup



Comparing with Fine-Tuned Models

Language	Best LLM	F1-Score	Fine-Tuned	F1-Score	Δ Val	Δ Best LLM
Danish	GPT-4.5	0.679	Babel Machine	0.925	+0.065	+0.246
Dutch	o1	0.724	Babel Machine	0.906	+0.066	+0.182
English	o1	0.706	Babel Machine ft-XLM-RoBERTa	0.869 0.810	-0.031 -0.009	+0.163 +0.104
French	o1	0.714	Babel Machine	0.821	-0.029	+0.107
Hungarian	GPT.4-5	0.672	Babel Machine	0.751	-0.099	+0.079
Italian	o1	0.675	Babel Machine	0.930	+0.120	+0.255
Portuguese	o1	0.651	Babel Machine	0.867	-0.063	+0.216
Spanish	o4-mini	0.756	Babel Machine	0.916	+0.066	+0.160

Note. All estimates are weighted F1-scores obtained on our fixed held-out test set. The columns Δ Val and Δ Best LLM indicate: (i) the change relative to the best result on the model's own validation set; and (ii) the change relative to the strongest zero-shot LLM, respectively.

How much data leakage is present here? Probably considerable, so that the results may be inflated. However, the potential advantages of fine-tuning still seem to outweigh in-context learning, even for BERT-like models.

Key Findings Meta-Analysis Policy Agenda

+8.9%

**Reasoning
Advantage**

Chain-of-thought capabilities boost F1-score by almost 9 points

-7.8%

**Reproducibility
Penalty**

Deterministic deployment reduces F1-score by about 8 points, but ensures consistency

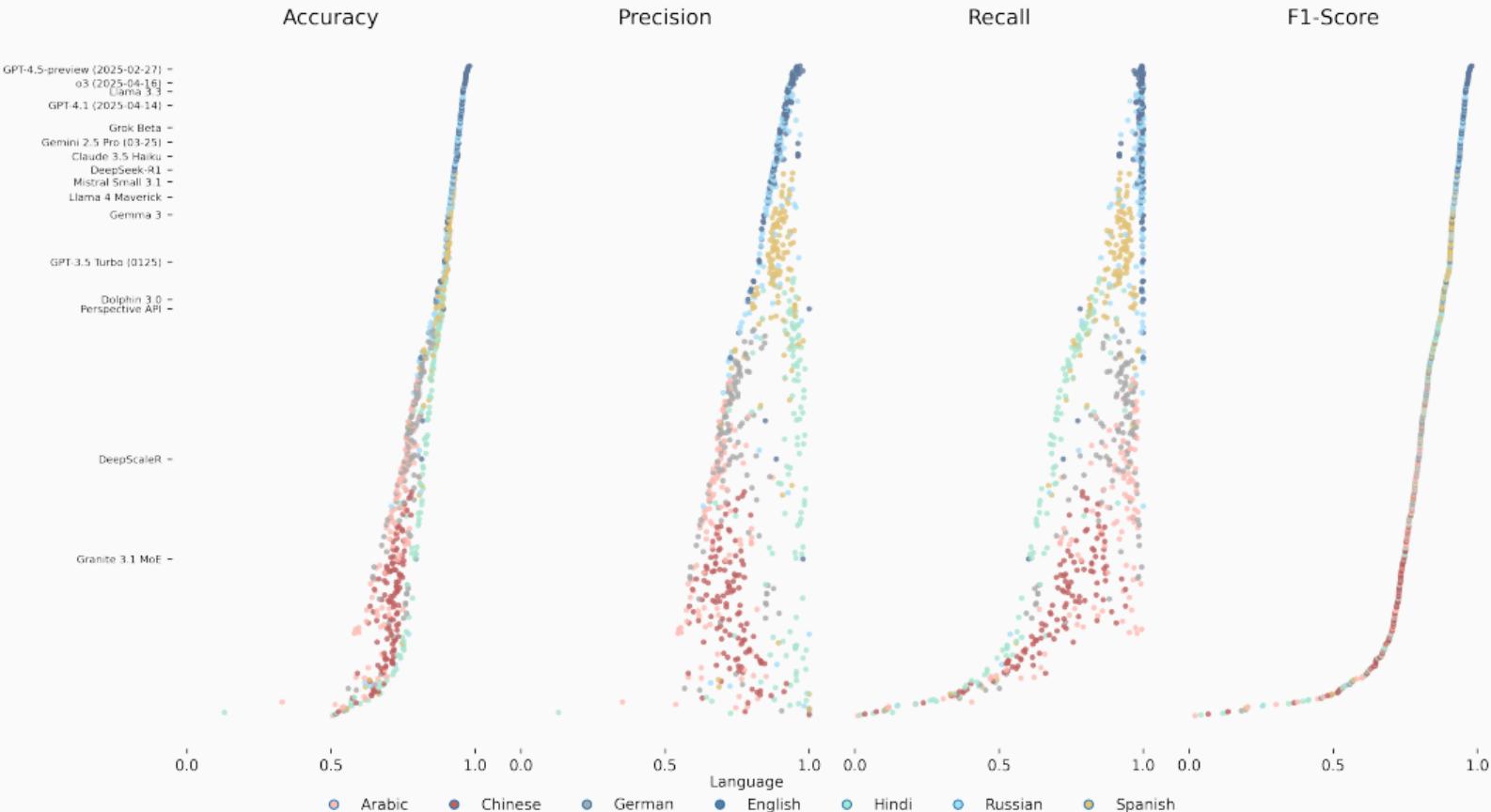
$\geq 10\%$

**Fine-Tuning
Advantage**

Supervised transformers outperform zero-shot LLMs by 10+ points

We did not find  openness penalty

Benchmarking Toxicity



Top Performing Models

Arabic

0.83 F1-Score

o3 · o1 · GPT-4.5

Chinese

0.78 F1-Score

GPT-4.1 · GPT-4o · Grok 3

English

0.98 F1-Score

Granite 3.2 ·
OLMo 2 · GPT-4.5

German

0.86 F1-Score

GPT-4.1 · o3 · GPT-4.5

Hindi

0.89 F1-Score

Mistral Saba · Mistral
Medium 3 · Gemma 2

Russian

0.96 F1-Score

o3 · GPT-4.1 nano
· Claude 4 Opus

Spanish

0.93 F1-Score

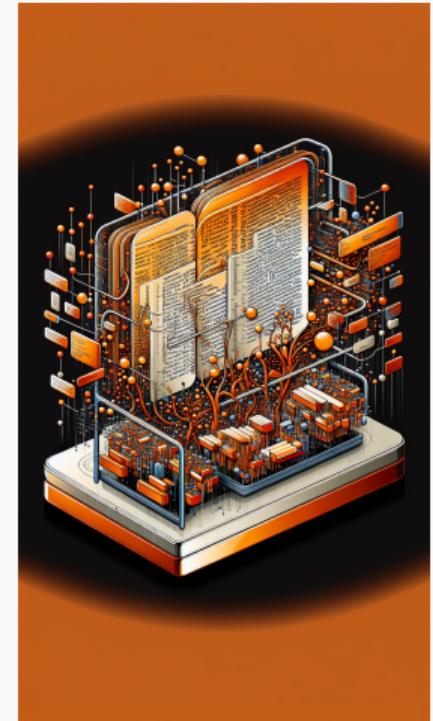
GPT-4.5 · o3 · GPT-4.1

Comparison to Perspective API

G Perspective API performance. Competitive accuracy but abysmally low recall scores, resulting in low F1-scores across languages.

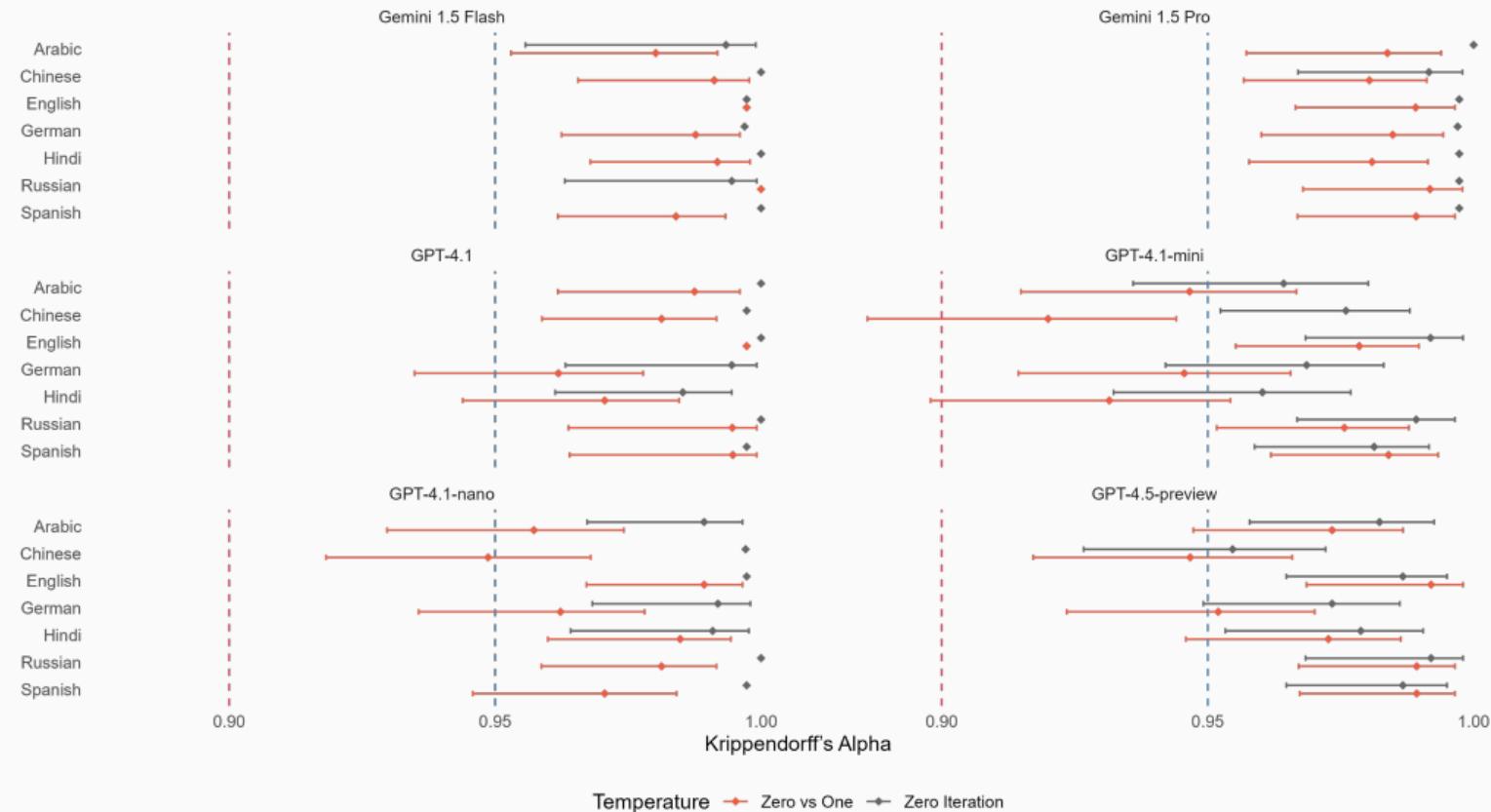
📊 Performance gap. Even weakest LLMs outperform Perspective by about 0.20 points in F1-score for Arabic and Chinese.

💻 Technological shift. Highlights transition from classical transformer pipelines to modern LLMs for toxicity classification.



Artwork by DALL-E 3 model

Temperature Experiments



Key Findings Meta-Analysis Toxicity

+7.7%

**Resource
Advantage**

High-resource languages enjoy a 7.7 points F1 advantage

-9.8%

**Size
Penalty**

Small models ($\leq 7B$) fall 9.8 points short of larger LLMs baseline

99%

**Annotation
Reliability**

Modern private LLMs exhibit outstanding reliability even w/temperature variations

We did not find openness or overthinking penalties

Methodological Takeaways

Evidence-Based Recommendations

- **Reasoning advantage.** Chain-of-thought capabilities enhance performance in policy agenda classification, but do not yield gains in toxicity classification.
- **Openness penalty inconclusive.** Slight lower performance than proprietary models, descriptively. There is no significant effect in meta-analyses. However, deterministic deployment reduces F1-scores in policy agenda classification.
- **BERT family pays off in complex tasks.** Supervised transformers outperform zero-shot LLMs by +10 points in policy classification. Even the weakest LLMs in toxicity outperformed off-the-shelf options.
- **Performance vs reproducibility.** Switching from deterministic to stochastic decoding rarely alters more than 1% of labels on average in toxicity.

This work has been supported by the Universidad Diego Portales (Inserción/Enlace Fund 2025), Agencia Nacional de Investigación y Desarrollo (ANID/PFCHA/72200340), Society for Latin American Studies (Postgraduate and Postdoctoral Award), St. Hilda's College, University of Oxford (Muriel Wise Fund), and OpenAI (Academic Programme).



Thank you very much!

Do you have any questions?

bastian.gonzalez.b@mail_udp.cl

bgonzalezbustamante.com